ASSIGNMENT-105

SP20-BCS-123
SYEDA NOOR ZEHRA NAQVI
GROUP IV

⇒ S_1 "sunshine state enjoy sunshine"
⇒ S_2 "brown fox jump high, brown fox run"
⇒ S_3 "sunshine state fox run fast"

⇒ Dictionary of vocabulary = {'brown', 'enjoy', 'fast', 'fox', 'high', 'jump', 'run', 'state', 'sunshine'}

Bag of words

	brown	enjoy	fast	fox	high	jump	run	state	sunshine	Total
S1	0	1	0	0	0	0	0	1	2	4
S2	2	0	0	2	1	1	1	0	0	7
S3	0	0	1	1	0	0	1	1	1	5

Vector S_1 : [0 1 0 0 0 0 0 1 2]
Vector S_2 : [2 0 0 2 1 1 1 0 0]
Vector S_3 : [0 0 1 1 0 0 1 1 1]

Term Frequency

TF	brown	enjoy	fast	fox	high	jump	run	state	sunshine
S1	0	0.250	0	0	0	0	0	0.250	0.500
S2	0.286	0	0	0.286	0.143	0.143	0.143	0	0
S3	0	0	0.200	0.200	0	0	0.200	0.200	0.200

IDF

= log(number of documents / no of documents with term)

idf	brown	enjoy	fast	fox	high	jump	run	state	sunshine
	0.477	0.477	0.477	0.176	0.477	0.477	0.176	0.176	0.176

TF-IDF

	S1	S2	S3
brown	0	0.136	0
enjoy	0.119	0	0
fast	0	0	0.095
fox	0	0.050	0.035
high	0	0.068	0
jump	0	0.068	0
run	0	0.025	0.035
state	0.044	0	0.035
sunshine	0.088	0	0.035

Cosine Similarity $= \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \times |\vec{y}|}$

Vector S1 = [0 1 0 0 0 0 0 1 2]

Vector S3 = [0 0 1 1 0 0 1 1 1]

$|S1| = \sqrt{1^2 + 1^2 + 2^2}$

$= \sqrt{6}$

$|S_3| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2}$

$= \sqrt{5}$

$S1 \cdot S_3 = (0\times0 + 1\times0 + 0\times1 + 0\times1 + 0\times0 + 0\times0 + 0\times1 + 1\times1 + 2\times1)$

$= 0+0+0+0+0+0+0+1+2$

$= 3$

Cosine similarity $= \frac{3}{\sqrt{6} \times \sqrt{5}}$

$= 0.5477$ Answer